公開

（事後評価・研究者用）

研　究　報　告　書

「空間データモデリングによるニューロンデータ検索の高速化」

研究期間：平成 30 年 11 月～平成 31 年 3 月
研究者番号：５０１３２
研 究 者：天方 大地

1．研究のねらい

ニューロサイエンス分野には，ニューロンのモデリングや神経回路の形成メカニズムの解明など，多くの問題が存在する．過去は実験室によるアナログな解析が主流であったが，コンピュータサイエンスの発展により，シミュレーション等による解析が可能となってきている．また，ニューロンの構造や脳を構成するニューロンの数が明らかになるにつれ，現実的なモデルを考慮した際の計算時間が問題となる．例えば，ニューロン間のインタラクション（シナプス結合を成すニューロンのペア）を計算する場合，単純にはニューロンの数の自乗の組み合わせが考えられるため，計算時間が爆発的に増加する．一方，科学的知見を得るためには，システムに対してインタラクトに問い合わせ（高速検索）が可能であることが理想であり，ニューロンの数に対してスケーラビリティを保有する検索アルゴリズムが要求される．

ニューロン間の信号処理やシナプス結合が起きる条件として，3次元空間における近接性が挙げられる．つまり，ニューロンの分析をより高度かつ高速に行うために空間データベース技術が必須である．空間データベースに関する研究は盛んに行われているが，既存研究では一つデータは一つの点と想定している．一方，ニューロンは単一の点ではなく，（離散化することで）大量の点から成る．ここで，ニューロン間の結合は axon と dendrite と呼ばれる部位間でのみ行われるという制約がある．そのため，ニューロンデータを取り扱う場合，既存技術の適応は困難であり，新たな技術開発が必要となる．本研究では，上記のようなデータを想定し，重要な役割を果たすニューロンの検索技術の開発を目的とする（ただし，アルゴリズムの出力を用いた実証実験等は本研究の対象外とする）．

2．研究成果

（1）概要

本研究では，1つのデータが複数の点で構成されているものに焦点を当てた．これは，ニューロンのデジタルデータに由来する．また，他のデータと最もインタラクションをとっているものを重要なデータと想定・定義し，そういったオブジェクトを検索するオペレータを設計した．

具体的には，ユーザが閾値 r を指定したとき，2 つのデータ間に距離が r 以内となる点のペアが存在する場合，それらのデータにはインタラクションがあると定義する．このとき，あるデータの重要度（スコア）は，インタラクションしているデータの数となる．このオペレータは，最もスコアが大きいデータを出力する．この問題を解く単純なアルゴリズムは，全ての点間の距離を計算することで，全てのデータのスコアを計算するものである．データの数を n，1つのデータに含まれる点の数の平均を m としたとき，このアルゴリズムの時間計算量は，$O(n^2m^2)$となり，データ数または点の数が大きいデータ集合に対応できない．また，オンライン

Advanced Information and Communication Technology for Innovation ACT-i

時間で $O(n\log n)$となるアルゴリズムが存在することも証明したが，このアルゴリズムは空間計算量が $O(n^2)$で，オフライン処理の計算量が $O(n^2(m\log m+\log n))$となり，実践的ではない．

そこで，実践的かつ効率的なアルゴリズムを設計した．このアルゴリズムは，新たなデータ構造である BlGrid (a hybrid index of compressed Bitset, Inverted list, and spatial Grid) に基づいて動作する．本問題のボトルネックは，データのスコアを計算する処理であり，高速に解を出力するためには，スコアを計算するデータの数を削減することが重要である．提案アルゴリズムでは，BlGrid を用いてスコアの下界値と上界値を計算することにより，解になり得ないデータをフィルタリングし，フィルタされなかったデータのみ正確なスコアを計算する．また，過去の検索の結果を利用した高速化，および，マルチコアを用いた高速化技術も開発した．これらの成果は，5-(1)-1 および 5-(1)-2 で発表した（する）論文にまとめられている．

（2）詳細

第1ステップ：重要なデータ（ニューロン）を検索するオペレータの設計．これは，インタラクションという概念を用いてデータの重要度・スコアを定義し，問題を定式化することによって達成した．（ユーザが閾値 r を指定したとき，2 つのデータ間に距離が r 以内となる点のペアが存在する場合，それらのデータにはインタラクションがある．）

また，この問題は，離散数学によってセマンティックを解釈できる．データのインタラクション関係をグラフで表す（図 1）．データを頂点，データ間にインタラクションがあれば，それらの頂点に辺があるとすると，頂点の次数はスコアを表し，オペレータが出力するデータは，次数中心性が最大のものとなり，ネットワークのハブを示す（黄色の頂点）．

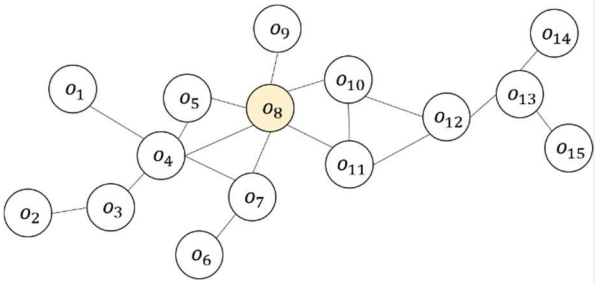

図 1：インタラクション関係のグラフ化

第２ステップ：問題の難しさ（挑戦度）の解析．上記の問題（オペレータ）は，非常にシンプルであるものの，効率的な解法は自明ではない．まず，解を出力するための単純なアルゴリズムは非効率的であることを示す．以降，データの数を n，１つのデータに含まれる点の数の平均を m とする．あるデータのスコアを計算する単純なアプローチは，自身の全ての点と，他の全てのデータの全ての点に対して距離計算を実行するものである．そのため，１つのデータのスコアの計算には $O(nm^2)$時間かかり，全てのデータのスコアを計算する時間は $O(n^2m^2)$となってしまう．この結果から，単純な方法はデータ数が多い場合に適切でないことが分かる．

次に，オンライン時間が $O(n\log n)$となるアルゴリズムが存在することを証明する．まず，全てのデータが他の全てのデータに対して最近傍ペアを計算し，その距離を昇順

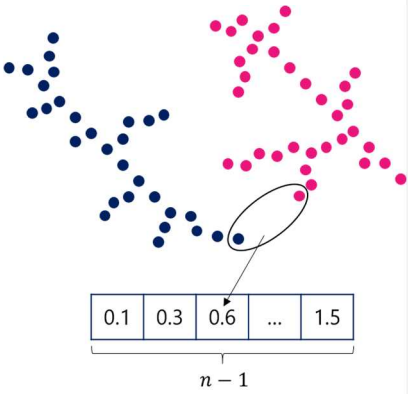

図 2：最近傍ペアの距離を格納する配列の例

に配列に格納する(図2)．この構造を用いると，r が与えれたとき，あるデータのスコアは，自身の配列において距離が r 以内となる要素の数となり，この検索は二分探索により $O(\log n)$で済む．そのため，全てのデータのスコアの計算には高々$O(n\log n)$時間しかかからない．しかし，このアルゴリズムには大きな欠点が2つ存在する．1つ目は，空間計算量が $O(n^2)$となり，データ数が大きい場合に対応できない．2つ目は，データ構造の構築に $O(n^2(m\log m+\log n))$時間かかってしまい，実際の解析を行えるようになるまで非常に時間がかかってしまう(実験では8時間を超えた)．

以上により，本問題の難しさを理論的に解析できた．この事実を踏まえると，実践的な効率性を持つアルゴリズムが必要となることが分かる．

第3ステップ：実践的かつ効率的なアルゴリズムの設計．本問題のボトルネックは，正確なスコアの計算処理である．出力されるデータ数は1つであることと，大半のデータのスコアは解に程遠い値であることを考慮すると，正確なスコアの計算を可能な限り削減することよって解を高速に出力できることが分かる．そのため，スコアの幅(下界値と上界値)を低コストで計算することより，解になり得ないデータをフィルタリングする．

このアイデアを実装するため，データのスコアの下界値と上界値を効率的に計算できるデータ構造である BIGrid (a hybrid index of compressed Bitset, Inverted list, and spatial Grid)を設計した(図3)．これは，グリッドの各セルに圧縮ビットセットや転置ファイルを持たせた構造である．スモールグリッドはスコアの下界値の計算に使われ，ラージグリッドはスコアの上界値および正確な計算に使われる．スコアの下界値と上界値の計算速度は正確なスコアの計算よりも圧倒的に早く，また，解になりえないデータも大量に発見できることから，処理の高速化を実現できた．

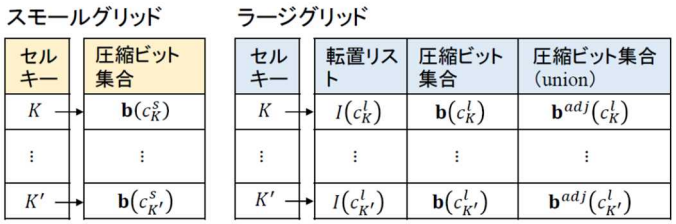

図 3：BIGrid の構造

さらに，過去の検索において，同じような閾値 r'が指定されていた場合，その検索で計算された結果を用いることによって処理を高速化する仕組みを設計した．これにより，上界値の計算処理を高速化できた．

加えて，マルチコアによる並列処理を可能とするようにアルゴリズムを拡張した．並列処理を行う際に重要なのは各コアの処理量を均一化することである．提案アルゴリズムは，BIGrid の構築，下界値の計算，上界値の計算，正確なスコアの計算の4ステップからなるが，それぞれコストモデルをたて，各コアのコストがほぼ均一になるようにデータ，または点を割り当てるアルゴリズムを設計した．その結果，コア数を増やすごとに処理時間が短縮されることを確認した．

3. 今後の展開

1つのデータが複数の点によって構成されるアプリケーションは，ニューロサイエンスに限らない．例えば，トラジェクトリデータベースやポイントクラウドも同様の構成を持つ．この観点から，本

研究はニューロサイエンスに限らず幅広いアプリケーションに利用できる分析ツールを提供していると考えられ，重要性が増す可能性がある．また，今回は次数中心性にのみ取り組んだが，次数中心性が高いデータが重要とは言い切れない場合もあり，他の中心性をテストする必要がある場合も考えられる．本研究の提案アルゴリズムは次数中心性にのみ対応しているため，他の中心性を取り扱うことはオープンプロブレムである．今後は，本研究を皮切りに関連した問題に取り組む研究が増加することを期待する．

4. 自己評価

- 当初予定していた研究目的は概ね達成できた．本研究では，次数中心性に焦点を当てたが，中心性の定義は他にもいくつか存在する．他の中心性に関するアルゴリズムの設計を行うまでに至ればより良かったように思う．
- 研究は基本的に個人で実施し，サイトビジットでのアドバイスをもとに研究をより良い方向に進めることができた．また，国際会議への参加を通して関連研究の動向を探り，使える技術を参考にすることもできた．また，英語論文執筆の際は，英文校閲が非常に役に立った．
- 本研究の成果はデータベース分野でのトップ会議に採択され，今後多くの研究者の目に止まることが予想される．そのため，さらなる効率化やモデルの向上･拡張，および産業界への応用が期待でき，多くのアプリケーションで利用されるまでに昇華されるものと信じている．
- 本研究では，空間データベースにおける新たなモデル上での検索オペレータとその効率的な解法を提案した．これまでに1つのデータを点集合して考えたオペレータとそのアプリケーションについては議論されていなかったことから，本研究の独創性を見て取れる．また，本研究の挑戦性については，計算時間の面から把握でき，2-(2)を参照されたい．

5. 主な研究成果リスト

(1) 論文（原著論文）発表

1. 天方 大地，原 隆浩，空間データベースにおけるインタラクティブオブジェクトの高速検索，第 11 回データ工学と情報マネジメントに関するフォーラム（DEIM フォーラム 2019），2019 年 3 月．
2. Daichi Amagata and Takahiro Hara, Identifying the Most Interactive Object in Spatial Databases, Proceeding of the IEEE International Conference on Data Engineering (ICDE), April, 2019 (to appear).

(2) 特許出願

研究期間累積件数：0 件

(3) その他の成果（主要な学会発表、受賞、著作物、プレスリリース等）

- (1)-2 の国際会議はデータベース分野におけるトップ会議であり，フルペーパー採択された．